全国第七届研究生数学建模竞赛

题 目 基于神经网络 MIV 值分析的肿瘤基因信息提取

摘 要：

本文主要运用统计学及数据挖掘相关知识，以结肠癌基因表达图谱为研究对象，综合运用 GB 指数、BP 神经网络、小波变换和贝叶斯等方法对问题给出求解的过程和结果。

问题一采用 GB 综合指数对无关基因进行筛选。首先计算各个基因的 *Gini* 指数和 Bhattacharyya 距离，其次合理定位阈值，分别在 *Gini* 指数排序和 Bhattacharyya 距离排序中选择 300 个优势位置的基因作为备用基因，最后选择这两组备用基因的交集作为信息基因，共 114 个。从而降低了基因维度。

问题二结合已有文献，本着创新性和有效性的思想进行基因信息提取。

- 首先利用基因间的强相关性进行初步冗余基因的剔除，得到了五组特征基因组；利用 BP 神经网络对这五组基因组进行错判数计算，选取错判率最低、基因子集中基因数量最少的基因特征组作为下一步研究基因子集；
- 其次利用平均影响值（MIV）方法来进行筛选基因，通过计算一个基因组合中每个基因的 MIV 值，每次剔除 MIV 绝对值处于后 10%的基因进行基因子集的确定；
- 最后利用 BP 神经网络来进行 22 个基因子集的错判数计算，最终确定含有 12 个基因的子集为最优基因组合（M85079，T62947，R39209，R84411，T54303，M82919，H43887，X12671，H08393，M26383，R36977，R87126）。

问题三将每组基因表达值看做基因信号，运用小波转换法对基因数据进行去噪，建立去噪模型。运用 MATLAB 小波工具箱对基因表达数据进行去噪处理，并运用去噪后的基因数据重新进行基因的分类、特征基因的提取。通过对比发现，去噪后的数据进行基因分类时能保留 61 个基因，比原始基因数据少 53 个，特征基因提取为 8 个。

问题四针对给定的若干信息基因，利用聚类分析原理和 Bayes 估计，通过模型建立给出探索其它未知信息基因的思想。

关键字：基因表达谱；GB 指数；MIV；BP 神经网络；小波变换；贝叶斯

参赛队号 10255012

队员姓名 王紫微 叶奇旺 徐信诚

参赛密码 _____

（由组委会填写）

中山大学承办

目 录

1. 问题重述.....1
2. 问题分析.....1
   2.1. 问题一.....1
   2.2. 问题二.....1
   2.3. 问题三.....2
   2.4. 问题四.....2
3. 基本假设.....2
4. 符号说明.....2
5. 模型建立及求解.....3
   5.1. 问题一：基因筛选.....3
      5.1.1. 实验数据分析.....3
      5.1.2. 信息基因的提取.....4
         （1）计算 Gini 指数.....4
         （2）计算 Bhattacharyya 距离.....6
         （3）GB 综合指标选取信息基因.....8
   5.2. 问题二：分类信息基因确定.....9
      5.2.1. 解题思路解析.....9
      5.2.2. 解题方法理论基础.....9
      5.2.3. 模型建立与求解.....10
         （1）特征基因的选取.....10
         （2）强相关性冗余基因的剔除.....10
         （3）基于 MIV 值的分类特征子集的选取.....11
         （4）特征基因组合的检验和比较.....13
   5.3. 问题三：基因信息去噪.....15
      5.3.1. 解题方法及思路分析.....15
      5.3.2. 基因信号去噪.....16
         （1）信号分解.....17
         （2）给小波细节系数加阈值.....18

（3）信号重建 19
5.3.3. 过滤后基因信号进行分类 21
（1）Gini 指标数据对比 21
（2）Bhattacharyya 距离对比 21
（3）基因提取 22
5.4. 问题四：信息基因求解的数学模型 23
5.4.1. 解题方法及假设 23
5.4.2. 贝叶斯分析 23
（1）Bayes 公式 23
（2）基于 Bayes 方法的聚类算法 24
5.4.3. 调用聚类算法 25
（1）确定初始聚类中心 25
（2）确定初始聚类种子及聚类中心点 25
（3）分类 25
（4）结果分析 25
6. 模型的评价与改进 26
6.1. 模型优点 26
6.2. 模型缺点 26
6.3. 模型的改进 27
参考文献 27

1. 问题重述

近年来，随着 DNA 微阵列（DNA Microarray）技术的发展，人类逐渐能够通过基因转译的基本原则检验得到人体基因的表达图谱。怎样从 DNA 芯片所测量的成千上万个基因中找出决定特定类别的一组“信息基因”，成为学者们在识别肿瘤类型进而研制抗癌药物过程中特别关注的问题。

问题一需要从附件给出的 2000 个基因中运用合理的分类方法剔除“无关基因”，得到较合理的“信息基因”供进一步分析。

问题二在第一题取得“信息基因”的基础上进一步分类筛选，得到供研究人员合理判断肿瘤的基因标签。

第三题要求通过建立噪声模型，对基因表达数据进行去噪处理，并分析去噪后的基因数据能否对确定基因标签产生积极的影响。

问题四要求小组成员在已知若干个信息基因的情况下，建立基因“标签”的数学模型，加入与癌症关系密切的信息基因。

2. 问题分析

2.1. 问题一

问题一中，主要考虑剔除基因表达谱中表达水平在两种样本（正常人和结肠癌患者）中都非常接近的基因。这些基因的均值和方差在各个样本中都无明显区别，不一定会对基因判别产生积极影响，反而可能增加信息基因搜索的计算复杂度。因此，通过剔除“无关基因”，能够大大提高基因分类提取的效率。

剔除无关基因可以运用多种方法实现。从“信噪比”指标筛选基因中的“信息基因”[1]，到 Bhattacharyya 距离、欧氏距离判定无用基因，这些方法虽取得了一定的成果，但同时存在各种缺陷。本文考虑运用 GB 综合指标结合 *Gini* 指数和 Bhattacharyya 距离两个因素对基因进行分类。

2.2. 问题二

肿瘤特征基因选取的目的在于从原始基因集合中提取出一组最能反映样本分类特性的基因以准确地刻画出事物的分类模型，从而为最终确定肿瘤分类与分型的基因标记物提供可靠线索。该特征基因集合应包含尽可能完整的样本分类信息，即不丢失原始基因集合中所蕴含的样本分类信息，做到对样本的准确分类。为了达到利用基因的表达水平进行肿瘤识别的目的，可以采用经过第一问处理后得到的 114 个基因作为样本特征分类器的输入。

从目前对该类型问题的研究文献中可以看出，刘全金、李颖新和阮晓钢[2]对肿瘤分类特征基因的选取采取多种方法实验均取得了理想的效果，分类的错误率相对较低。相关研究主要通过以下三个步骤对肿瘤基因进行选取：

- 无关基因的剔除——多数使用“信噪比”、Bhattacharyya 距离等方法

- 分类特征基因的提取——在该步骤中方法的使用可谓是百花齐放，主要方法有：BP 神经网络的灵敏度分析、顺序浮动搜索算法、统计方法等，这几种方法在该类问题中的应用得到了很好的检验，同时也取得了比较满意的效果。
- 分类效果的检验——主要利用 “留一交叉检验分类法”和“独立测试分类法”，在工具上的选择则比较多：支持向量机、遗传算法、BP 神经网络等，这几种方法的使用效果差距不是很大，主要在于使用者的偏好。

本组进行讨论并最终选用 BP 神经网络进行基因筛选。

2.3. 问题三

基因信息在芯片制作、基因表达试验过程中会不可避免的产生噪音。在本问中，小组成员综合比较均值去噪、中值去噪等方法，针对基因数据样本少、维数多的特点本文采用小波变换方法进行去噪处理，并将去早后的数据进行归类，观察去噪处理对基因筛选的作用。

2.4. 问题四

问题四在第二问的基础上加入若干个已知的信息基因，判断已知部分信息基因的加入是否会对甄别其它未知信息基因产生影响，这就需要我们利用好已给的信息基因和信息基因的概率这两个有用的信息来建立数学模型，本文从贝叶斯分析利用先验概率和质心法聚类这两部分来建立数学模型。

3. 基本假设

假设 1，该题的基因数据相对真实有效，可用于数据挖掘分析；

假设 2，假设各个基因之间的交互作用小，可以将单个基因当做独立影响指标来进行分析；

假设 3，基因数据样本来自于同一时间段，不考虑由时间因素引起的数据误差。

4. 符号说明

问题一：n_{ij}——基因i在第j个样本中的表现值

$Gini(k)$——类别k的 Gini 指标

P_{ij}——基因i在样本j中对类别k的相对频率

B——Bhattacharyya 距离

μ_1——正常人基因表达数的均值

μ_2——结肠癌患者基因表达数的均值

σ_1——正常人基因表达数的方差

σ_2——结肠癌患者基因表达数的方差

G——基因总体

A_I——信息基因集合

A_N——无关信息集合

θ——Bhattacharyya 距离阈值

问题二：MIV——平均影响值

IV——影响值

F_i——i 维候选特征基因子集

X_{ij}——基因 i 在样本 j 中的表达值

X_{ij}'——基因 i 在样本 j 中的规划值

问题三：n_i——零均值的高斯白噪声

y_i____为含噪信号

x_i____为需要提取的有用信号

B——噪声方差

问题四：$P(A/B)$——B 事件发生条件下 A 事件发生的概率

5. 模型建立及求解

5.1. 问题一：基因筛选

针对问题一，需要考虑通过剔除“无关基因”来提高基因分类提取的效率。本文通过分析决定运用 *Gini* 指数与 Bhattacharyya 距离综合考虑构建 GB 指标对无关的基因数据进行筛选。

5.1.1. 实验数据分析

结肠癌基因芯片的数据中具有 40 个结肠癌的组织样本和 22 个正常组织样本，每个样本包含 2 000 个基因的表达数据。先对样本数据进行归一化处理，再将正常（normal）样本和结肠癌（cancer）样本按照接近 2:1 的比例随机地分配到训练集和测试集中。如下图所示，训练集有 40 个样本，测试集有 22 个样本。

训练集		测试集
Cancer 26	+	Cancer 14
Normal 14		Normal 8

图 1 基因表达谱实验数据集

5.1.2. 信息基因的提取

本文选用*Gini*指数作为评价基因分类信息含量多少的指标，将之与 Bhattacharyya 距离相结合构建 GB 综合指标。

*Gini*指数是数据挖掘中用来评价分类节点优良程度的常用指标。本文认为基因的*Gini*指数理论上可以体现出基因数据的不平等性对样本分类的贡献大小。*Gini*指数值越大，数据之间均等程度越高，基因的分类信息量越少；*Gini*指数值越小，数据的表达水平越不均等，基因的分类信息量越大。

基因的 Bhattacharyya 距离体现出基因在两个类别中分布均值的差异对样本分类的贡献和分布方差的不同对样本分类的贡献，Bhattacharyya 距离越大，该基因对样本的可分性就越好。

将*Gini*指标与 Bhattacharyya 距离结合进行判别可以同时从多个方面考察基因对分类贡献的大小。GB 指标的构建方法与计算过程如下：

- 计算各基因的*Gini*指数和 Bhattacharyya 距离
- 将基因按两个指标值分别进行排序，提取分类信息量较大的前 m 个信息，得到两组基因个数都为 m 的备用基因；
- 从两组备用基因中提取公共基因，提取时以 Bhattacharyya 距离排名顺序为主，*Gini*指数排名顺序为辅。提取的公共基因的顺序作为公共基因的 GB 综合指标排名

（1）计算 Gini 指数

计算*Gini*指数前首先要将基因表达值离散化。本文将每个基因表达值离散化到 0—20 这 21 个等级。设基因 g_i，表达水平的最大值和最小值分别为 $\max(i)$ 和 $\min(i)$，则基因 g_i 在第 j 个样本中的值 n_{ij} 离散化后得到的值 S_{ij} 为：

$$S_{ij} = INT\left[20 \times \frac{n_{ij} - \min(i)}{\max(i) - \min(i)} + 0.5\right] \quad i = 1,2,3.....2000 \ \ j = 1,2,3.....62 \qquad (1)$$

其中 $INT[]$ 表示取整，表达值离散化后，计算各基因的*Gini*指数值，基因 g_i 对类

别 k（1 表示正常人样本，2 表示结肠癌患者样本）的 *Gini* 指标为：

$$Gini(k) = 1 - \sum_{j=0}^{20} \left[p_{ij} \right]^2 \qquad (2)$$

P_{ij} 为基因 g_i 在样本 j 上对类别 k 的相对频率。当 *Gini*（k）为 0 时，即对类别 k 所有样本都属于同一级别，表示能得到最大的有用信息；当 k 处所有样本没有相同级别时，*Gini*（k）最大，表示能得到最小的有用信息。

基因 g_i 的 *Gini*（g_i）就是：

$$Gini_{split}(g_i) = \sum_{k=1}^{2} \frac{n_k}{n} Gini(k) \quad i = 1,2,3.....2000 \qquad (3)$$

其中，n 是样本总数；n_k 是第 k 类亚型的样本数，在本文中，k 只能取 1 或者 2。

根据公式（1）、（2）、（3）计算 2000 个基因的 *Gini* 值按升序排列，取值较小的前 m 个基因作为构建 GB 综合指数的备用基因。通过以上步骤计算的部分基因的 *Gini* 值见下表：

表 1 部分基因 *Gini* 表

GenBank Acc No	S_{ij}		Gini(1)	Gini(2)	Ginisplit	GB 值
	cancer9	cancer10				
T79595	10	9	0.847107	0.90375	0.883651026	0.04148752
T41094	6	9	0.892562	0.91625	0.907844575	0.024422141
K01144	12	7	0.904959	0.92375	0.917082111	0.021050758
M37721	11	4	0.909091	0.9225	0.917741935	0.000346151
L23823	3	12	0.909091	0.92	0.916129032	0.011404985
M28827	7	11	0.88843	0.92125	0.909604106	0.018255405
H15901	6	12	0.904959	0.91	0.908211144	0.004186389
X00588	2	1	0.904959	0.925	0.917888563	0.013737957
T72582	9	13	0.88843	0.91	0.902346041	0.003229891
T71207	1	8	0.88843	0.915	0.905571848	0.011038673
L06175	6	12	0.847107	0.885	0.871554252	0.011768144

L19956	1	7	0.896694	0.915	0.908504399	0.008331145
U12140	12	13	0.863636	0.90375	0.889516129	0.000355987
R77780	6	0	0.88843	0.91875	0.907991202	6.88433E-05
X03674	5	11	0.900826	0.91	0.906744868	0.021143508
T67897	3	10	0.884298	0.91	0.900879765	0.008050495
X01410	2	4	0.896694	0.8875	0.890762463	0.002692422
D14695	3	1	0.92562	0.9225	0.923607038	0.00442951
R71401	4	4	0.904959	0.8925	0.896920821	0.012098989
X55019	3	5	0.900826	0.91875	0.912390029	0.014878934
R07333	0	3	0.913223	0.92	0.917595308	0.007679409
M55683	2	10	0.896694	0.92375	0.91414956	0.008012708

各基因排序后的分布曲线如下图所示：

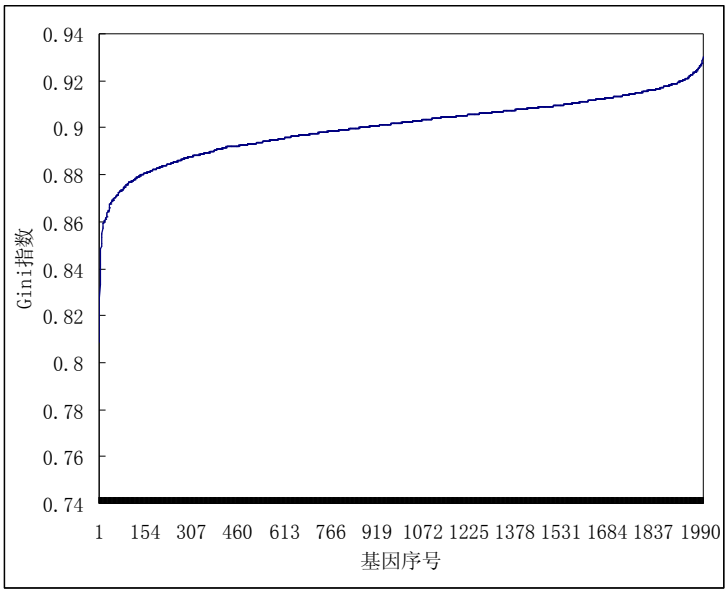


图 2 基因的 *Gini* 指数值升序排列分布曲线图

（2）计算 Bhattacharyya 距离

计算基因的 Bhattacharyya 距离的公式为：

$$B=\frac{(\mu_1-\mu_2)^2}{4(\sigma_1^{\,2}+\sigma_2^{\,2})}+\frac{1}{2}\ln\left(\frac{\sigma_1^{\,2}+\sigma_2^{\,2}}{2\sigma_1\sigma_2}\right) \qquad (4)$$

其中 μ、σ 分别表示基因的均值和标准差。运用 SPSS 统计软件将不同样本中的基因指标进行描述统计处理，得出正常人和结肠癌患者基因表达指标的均值和方差（具体数据见附件/GB 方法剔除无关基因.xls）。

运用公式（4）计算得出正常人基因与结肠癌患者基因的 Bhattacharyya 距离，通过频数分析得出不同巴氏距离值对应的基因数量如表 2 所示：

表 2 基因 Bhattacharyya 距离发布情况

Bhattacharyya 距离	基因个数	所占百分比
0～0.05	1571	78.55%
0.05～0.1	311	15.55%
0.1～0.15	74	3.70%
0.15～0.2	26	1.30%
0.2～0.25	12	0.60%
0.25～0.3	5	0.25%
0.35～0.35	1	0.05%

由于基因的 Bhattacharyya 距离越大，利用该基因的信息样本可分性就越好。因此 Bhattacharyya 距离大的基因数据理论上对于区分结肠癌基因起到更多的作用。

根据基因所含样本类别信息的多少，可以将基因分为“信息基因”和“无关基因”。用 A_I 表示“信息基因”，用 A_N 表示“无关基因”，则：

$$G \in \begin{cases} A_I & B(G) > \theta \\ A_N & B(G) \le \theta \end{cases} \qquad (2)$$

式（2）中，$B(G)$ 表示基因的 Bhattacharyya 距离，θ 表示指定的 Bhattacharyya 距离阈值。由表 2 数据可以看出，78.55%的基因 Bhattacharyya 距离在 0 到 0.05 个单位之间，剩下 21.45%的基因的 Bhattacharyya 距离分布在 0.05 到 0.35 之间。因此，可认为大部分的基因巴氏距离小于 0.05，这些基因在正常人和结肠癌患者两个类别中分布的均值和方差均无明显差异，因此可以将阈值 θ 定为 0.05，基因巴氏距离小于 0.05 的值可以作为无关基因剔除。

根据表 2 作出基因 Bhattacharyya 距离分布频数的直方图，见下图。

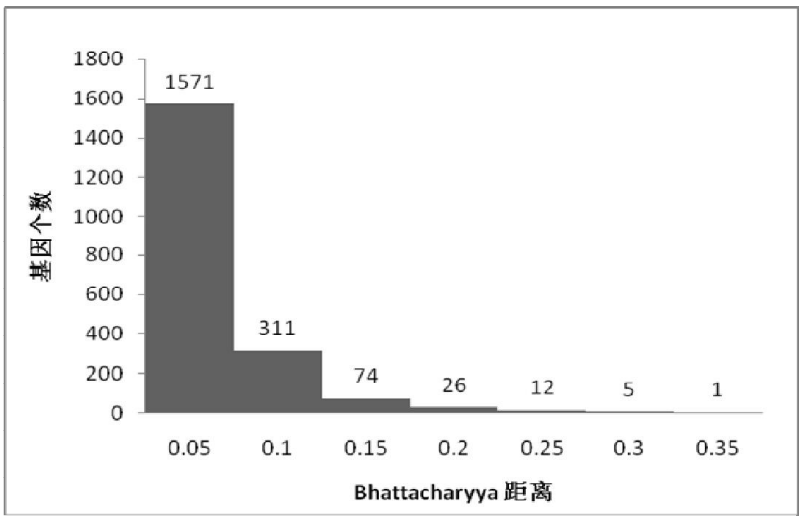


图 4 基因 Bhattacharyya 距离分布直方图

将 2000 个基因的 Bhattacharyya 距离值按降序排列，取值较大的前 m 个基因作为备用基因，降序排列后的 Bhattacharyya 距离分布曲线如下图所示。

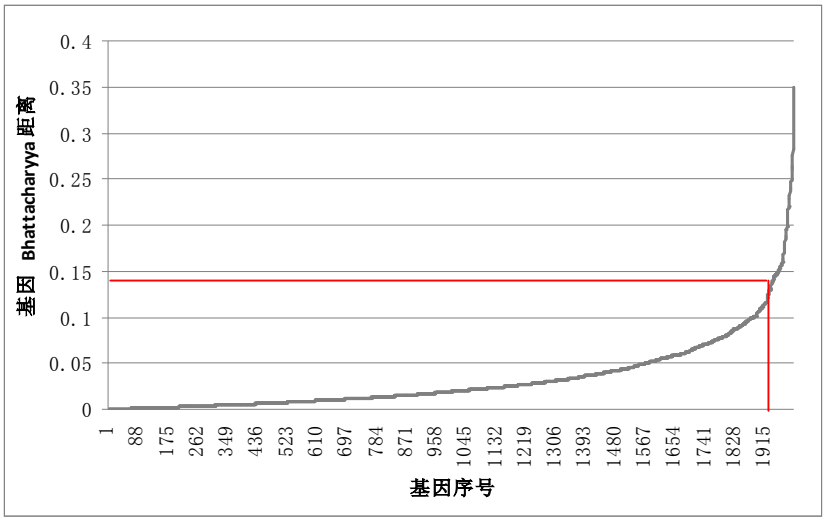


图 5 基因 Bhattacharyya 距离累计分布图

（3）GB 综合指标选取信息基因

从上面两种方法的分布曲线图中可以看出，只有小部分基因的分类信息量较高，其余基因的分类信息量都很低，因此，选取 m 为基因总量的 15%，其余基因作为无关基因剔除。保留的两组备用基因：*Gini* 指数排名靠前的 300 个基因和 Bhattacharyya 距离排名靠前的 300 个基因。提取两组备用基因中的公共基因，提取时按 Bhattacharyya 距离排名为主，*Gini* 指数排名为辅，公共基因的顺序作为相应的 GB 综合指标排名顺序。GB 综合指标排名越靠前的基因分类信息量越大。通过计算，在 300 个基因中共选出 **114** 个公共基因。

5.2. 问题二：分类信息基因确定

5.2.1. 解题思路解析

本文在进行问题分析和方法的选择时主要考虑两个方面：创新性和有效性，综合前人的研究成果和现有工具的优缺点，在进行分类特征基因的提取时选择了两种创新方法：强相关性去冗余基因和应用平均影响值（MIV，Mean Impact Value）方法结合神经网络的方法；在进行分类效果的检验时选择自组织竞争神经网络的方法进行分类。

5.2.2. 解题方法理论基础

（1） BP 神经网络

BP（Back Propagation）网络是 1986 年由 Rumelhart 和 McCelland 为首的科学家小组提出，是一种按误差逆传播算法训练的多层前馈网络，是目前应用最广泛的神经网络模型之一。BP 网络能学习和存贮大量的输入-输出模式映射关系，而无需事前揭示描述这种映射关系的数学方程。它的学习规则是使用最速下降法，通过反向传播来不断调整网络的权值和阈值，使网络的误差平方和最小。BP 神经网络模型拓扑结构包括输入层（input）、隐层(hide layer)和输出层(output layer)。网络结构见下图：

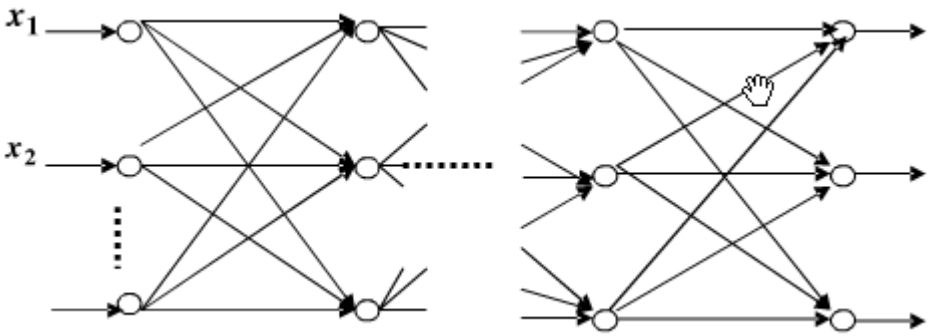


图 6　BP 网络结构

（2） MIV 值计算

Dombi 等人提出用 MIV 来反映神经网络中权重矩阵的变化情况，MIV 被认为是在神经网络中评价变量相关性最好的指标之一，也为解决此类问题开创了新思路。

本题选择 MIV 作为评价各个自变量对于因变量影响的重要性大小指标。MIV 是用于确定输入神经元对输出神经元影响大小的一个指标，其符号代表相关的方向，绝对值大小代表影响的相对重要性。具体计算过程，在网络训练终止后，将训练样本 P 中每一自变量特征在其原值的基础上分别加（减）10%（或自定义变化值）构成新的两个训练样本 P1 和 P2，将 P1 和 P2 分别作为仿真样本利用已建成的网络进行仿真，得到两个仿真结果 A1 和 A2，求出 A1 和 A2 的差值，即为变动该自变量后对输出产生的影响变化值（IV，Impact Value），最后将 IV 按观测列数平均得出该自变量对于应变量

——网络输出的 MIV。按照上面步骤依次算出各个自变量的 MIV 值，最后根据 MIV 绝对值的大小为各自变量排序，得到各自变量对网络输出影响相对重要性的位次表，从而判断出输入特征对于网络结果的影响程度，即实现了变量筛选。基于 BP 神经网络的 MIV 值的变量筛选流程见下：

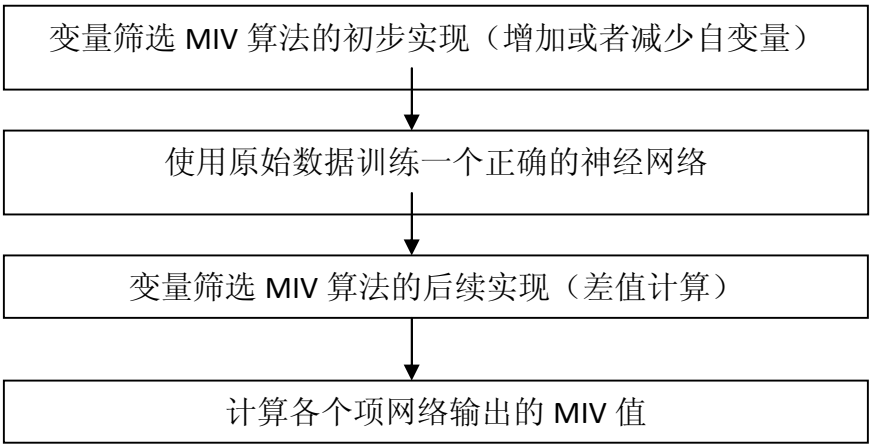


图 7　基于 BP 神经网络的特征基因选取的流程图

5.2.3.　模型建立与求解

（1）特征基因的选取

本问是在第一问的基础上再进行特征基因的选取，前面得到的 114 个基因汇总还可能存在冗余，这些冗余基因尽管也包含了样本分类信息，但其存在与否并不会影响整个分类特征基因集合的样本分类能力。在此，本题进一步研究特征基因的进一步筛选。

（2）强相关性冗余基因的剔除

从生物学角度分析，基因之间存在着调控和相互作用的关系,这在表达谱中反映为不同基因在表达水平上存在着一定程度上的相关性。据此可以进行冗余基因的初步排除。利用 SPSS 工具对得到的 114 个基因两两之间的相关系数，若其相关系数大于指定阈值，则认为两个基因是强相关的，排除二者中分类信息指数较小（GB 值较小）的那个基因，这使得排除冗余后的分类特征基因集合具有较大的分类信息指数。这一冗余剔除的过程在可以称为“两两冗余”分析。

利用 SPSS 来计算基因 g_i， g_i 之间的 Person 相关系数，具体计算公式如下：

$$Corr\_Coef(g_i,g_j)=\frac{\sum_{k=1}^{n}(x_{gik}-\overline{x}_{gi})(x_{gjk}-\overline{x}_{gj})}{\sqrt{\sum_{k=1}^{n}(x_{gik}-\overline{x}_{gi})^2\sum_{k=1}^{n}(x_{gjk}-\overline{x}_{gj})^2}} \qquad (5)$$

式中 x_{gik}， x_{gjk} 为基因 g_i,g_j 在训练集第 k 个样本中的表达水平值，$\overline{x}_{gi}$， $\overline{x}_{gj}$ 分别为 g_i,g_j 在训练集所有样本中表达水平的均值。该题通过选择几个相关系数的阈值来进

行分析。表 3 列出了在不用阈值条件下进行“两两冗余”分析后剩下的分类特征基因的数量，以及利用这些基因进行样本识别时对样本的分类能力。识别实验采用的是后面介绍的自组织竞争神经网络模型。

表 3 不同阈值条件下得到的分类特征基因的数量及其分类能力

不同阈值	1	0.9	0.85	0.8	0.75	0.725
剩余分类特征基因数量	114	83	46	30	17	10
分类错误数	2	2	3	5	5	6

由上表知，当阈值取 0.85 时,得到了能够对原始样本集中每个样本都能正确分类的最小分类特征数。这样，原来的 114 个分类特征基因经“两两冗余”分析，得到了 46 个分类特征基因，这 46 个基因组成的集合记为 S_I。

（3）基于 MIV 值的分类特征子集的选取

“两两冗余”分析考虑的是“单个”基因间的相互关系。然而基因间的调控和相互作用往往表现为“功能组合基因”的形式，即基因的功能与作用往往是“某些”基因的组合作为一个整体共同作用的结果。从“组合基因”的角度出发进行分类特征基因的选取更符合生物学实际。

基于 BP 网络分类器分析基因的 MIV 值是为了从整体考察基因集合对样本分类的影响，度量每个基因在群体中发挥的作用大小。这样，基因的 MIV 就可以看成是其影响判断肿瘤样本的重要性指标，可以从基因集合中去除影响较小的基因，保留影响大的基因，从而得到维数较小的候选特征基因集合。考虑到基因间存在着复杂的调控关系，一个去除影响较小的基因的有效策略是采用逐步剔除法，即首先去除具有最小 MIV 值的基因，然后对剩余基因整体再一次考察其中各个基因的 MIV 值，并根据重新计算得到的灵敏度去除其中灵敏度最小者，该过程反复进行，直到特征集合中基因全部被剔除为止。上述过程可描述如下：

（1）在训练集中用候选特征基因子集 F 构建并训练 BP 神经网络模型；

（2）计算 F 中各基因对输出函数的 MIV 值；

（3）对 F 中的基因按 MIV 值由大到小排序，去除 F 中具有最小 MIV 值的基因 g，得到新的候选特征基因子集：$F = F - \{g\}$；

（4）若 $F \neq \Phi$,则返回步骤（1）继续执行，否则退出。

在实际运行中，每次剔除占 F 中后 10 %的基因以加快算法运行速度。具体实现过程见下面。

利用筛选的 114 个基因进行 MIV 值的计算，通过 MATLAB 进行 MIV 值求解的结果如下（代码见附录二）：

神经网络的训练结果如下图：

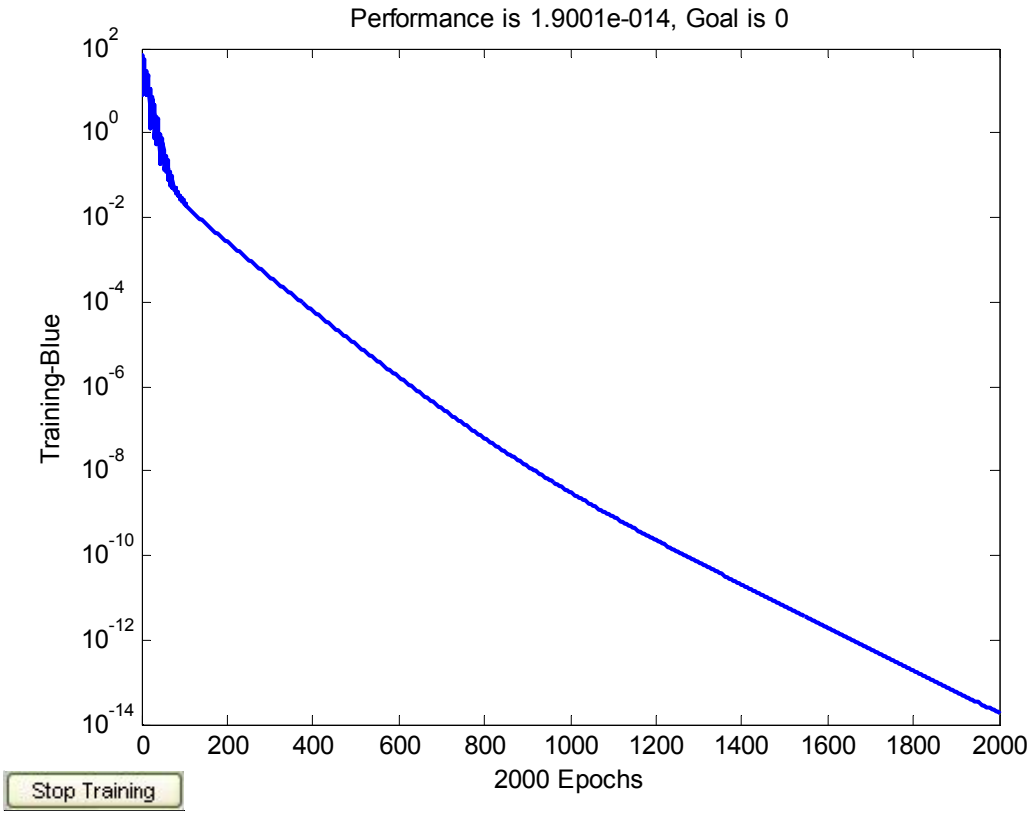


图 8 BP 神经网络学习训练的效果

从图 6 中可以看出，通过 2000 次的迭代学习，目标函数的收敛值为 1.9001e-014，已经相当接近 0 了。运算出来的 MIV 值见下表：

表 4　F114 的 MIV 值

基因名称	H87135	R62549	U02493	R34876	H77597	H25136	Y00345	R98842	M15841	U19969
MIV	-0.08	-0.42	1.10	0.04	0.28	-0.43	0.08	0.29	-0.01	-0.24
基因名称	D00860	H17434	X12369	T59878	T90280	T60778	L24203	L11706	L13738	X07290
MIV	-0.38	0.37	-0.07	0.71	0.68	-0.32	0.30	0.06	0.14	0.44
基因名称	T40454	Z25521	T95048	R75843	H11084	U09587	T79152	R97912	X53586	X68314
MIV	-0.05	-0.33	-0.46	-0.63	0.02	-0.78	0.27	-0.47	-0.51	0.11
基因名称	T96873	T51858	U25138	X15183	U32519	D21261	H05899	M35252	H61410	H20426
MIV	1.04	-1.32	-0.05	1.02	-0.47	-0.13	-1.05	-0.07	0.32	-1.41
基因名称	M34175	L41559	L09604	H40560	U22055	T58861	D38551	T52185	U21090	R54097
MIV	-0.83	0.02	0.83	-0.62	-0.06	1.41	0.77	-0.39	0.37	-0.03
基因名称	T61609	T51261	T83368	R08183	X70326	M85079	J02763	T51529	U17899	T48804
MIV	-0.54	-0.50	0.44	-0.90	1.02	1.55	0.01	-0.62	-0.40	0.82

基因名称	R42501	R10066	T57468	M36981	D31885	T62947	T92451	T51023	T86749	U26312
MIV	0.35	0.27	-0.31	1.39	-0.06	0.77	0.65	-0.92	0.50	0.80
基因名称	H55916	R64115	X86693	T95018	X56597	R39209	H23544	X55715	T47377	T57619
MIV	0.27	0.71	-0.52	0.93	0.54	-0.07	0.99	-0.14	-0.02	0.38
基因名称	T51571	H55758	M26481	T70062	H06524	X12466	X54942	R39209	T71025	M26697
MIV	0.23	1.52	-0.30	1.24	-0.12	0.66	0.87	-0.54	-0.30	0.16
基因名称	X14958	T86473	T56604	M76378	H40095	U09564	R84411	M36634	T54303	M76378
MIV	0.66	-0.73	-0.90	0.72	0.31	-0.75	-1.51	0.70	-0.21	-0.36
基因名称	M82919	H43887	X63629	Z50753	U30825	J05032	J02854	M76378	X12671	H08393
MIV	0.27	0.03	0.52	-0.20	-0.05	-1.66	0.03	-0.24	1.00	0.09
基因名称	M26383	M63391	R36977	R87126						
MIV	-0.12	0.82	0.24	-0.42						

再进行 MIV 值的排序和筛选，去掉 MIV 值的绝对值处于后 10%的基因，利用甚于的基因进行重复上边的工作，直到最后的基因数为 1 为止。这样可以得到以下候选特征基因子集：F_{46}，F_{40}，F_{35}，F_{32}，F_{29}，F_{26}，F_{23}，F_{20}，F_{18}，F_{16}，F_{14}，F_{12}，F_{10}，F_9，F_8，F_7，F_6，F_5，F_4，F_3，F_2，F_1（下标为候选特征基因子集的特征维数）。

（4）特征基因组合的检验和比较

此步采用的是“独立测试实验”方法来进行错判率的计算，首先利用训练集中的所有 62 个样本训练 BP 神经网络，然后利用该网络对测试中 22 个样本进行分类测试，并记录分类结果和错判率，重复该过程 10 次，计算错判率的均值，以此均值作为该基因组合的最终错判率。

下边以初始的 114 个基因组合为蓝本介绍相关的过程。

首先，将数据进行归一化处理

$$X_{ij}' = \frac{X_{ij}}{\sum_{j=1}^{n} X_{ij}} \qquad (6)$$

然后，把相关数据导入 Matlab 后，进行训练的结果如下：

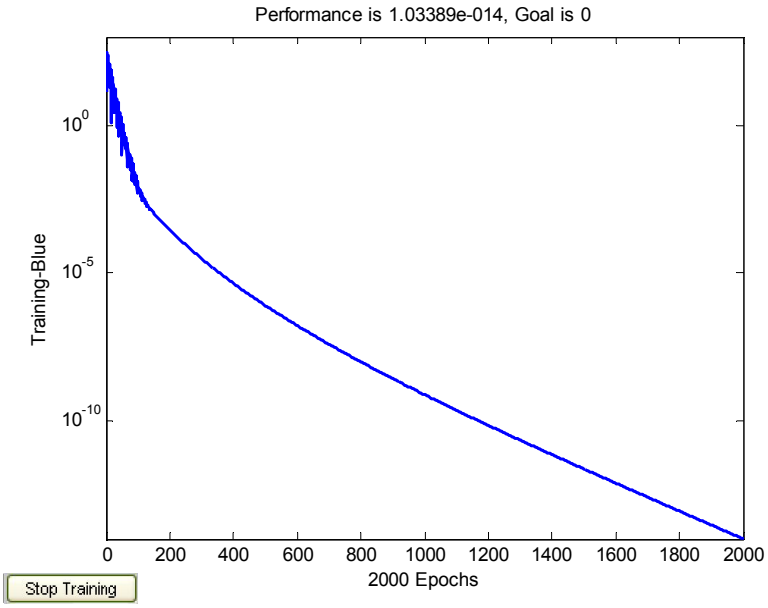


图 9　BP 神经网络学习训练的效果

从上图可以看出，训练的效果很好，收敛成功。

预测的结果见下表：

表 5　F114 基因组合的分类结果

编号	1	2	3	4	5	6	7	8	9	10	11	12
真实	1	1	1	1	1	1	1	1	2	2	2	2
分类	1	1	1	1	2	1	1	2	2	2	1	2
编号	13	14	16	16	17	18	19	20	21	22		
真实	2	2	2	2	2	2	2	2	2	2	**错分样本数：**	
分类	2	2	2	2	2	2	2	1	2	2	**4**	

重复上述步骤 10 次，即可得到该基因组合的 10 次数据，通过求分错率的平均数即可求得 F_{114} 的分错率。

通过以上步骤进行重复实验，对每种基因组合都进行预测和计算其分错率，最终可以得到下表。

表 6　各种基因子集的分错率统计

基因组合	F114	F83	F46	F40	F35	F32	F29	F26	F23	F20	F18	F16	F14
方法	GB 计算	两两冗余		BP 神经网络的 MIV 值分析									
错判数	4	3	3	2	4	2	4	2	2	2	2	2	2
基因组合	F12	F10	F9	F8	F7	F6	F5	F4	F3	F2	F1		
方法	BP 神经网络的 MIV 值分析												

错判率	2	3	4	4	3	5	5	6	6	5	7	

上表的基因子集的错判数见下边折线图：

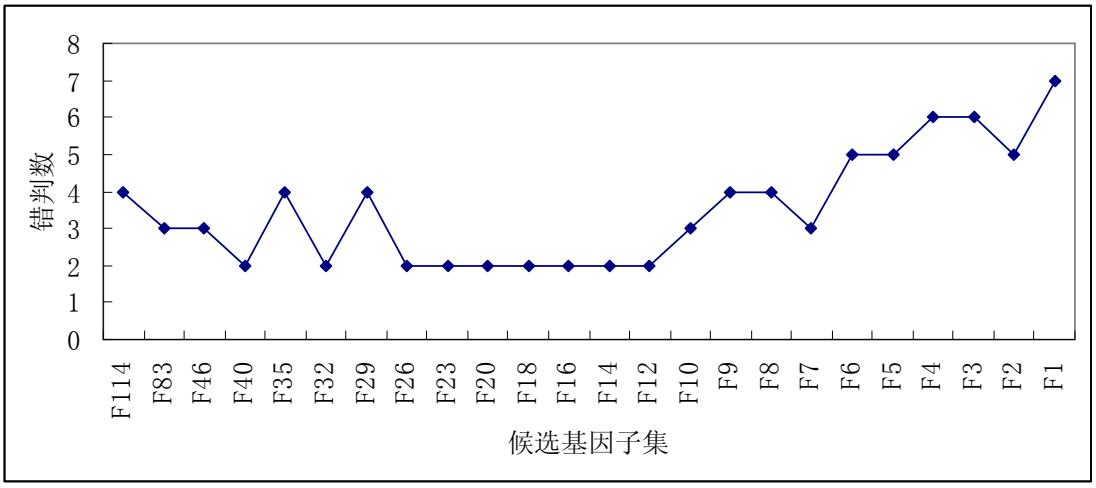


图 10　不同维数候选特征子集的分类能力

通过上表选取基因数量最少同时错判率最低基因组合，最终的基因子集为 F12，有以下基因：M85079，T62947，R39209，R84411，T54303，M82919，H43887，X12671，H08393，M26383，R36977，R87126。

将以上基因和 Guyon 学者[3]所筛选出来的基因“标签”进行对比，发现 H08393，T62947 这两个基因同时也出现在 Guyon 的研究结果中；与刘全金学者的研究成果[4]进行比较，发现 R87126，T62947，H08393 这三个基因也同时出现在其论文研究中。

5.3. 问题三：基因信息去噪

5.3.1.　解题方法及思路分析

基因表达数据在芯片制作、基因表达试验或者基因芯片表达图谱的采集转换过程中都会或多或少的引入噪音，这种噪声甚至会覆盖数据点。因此，利用合理的方法对基因数据进行去噪同样是一个至关重要的问题。

目前，均值滤波、中值滤波、数学形态学滤波和小波变换滤波等方法被广泛应用于基因芯片图像去噪研究。分析各种去噪方法的优缺点，本小组采用小波变换方法为基因进行去噪。去噪后重复问题一和问题二的步骤验证去噪过程是否能够为确定基因标签产生有力的影响。

小波变换作为一种强有力的信号分析工具，是 Morlet 于 20 世纪 80 年代末在分析地球物理信号时提出来的，它是泛函分析、傅立叶分析、样条理论、数值分析等多个学科相互交叉的结晶，是分析非平稳信号的有力工具，既可以分析信号的概貌，又可以分析信号的细节。在图像去噪领域，小波变换的许多优良性质使之得到了广泛的应用。

小波去噪的基本思想是通过小波变换，把信号分解为低频信号和高频信号，低频成分给出了信号的特征，高频成分则与噪音和扰动连在一起。将信号的高频成分去掉，

信号的基本特征仍然可以保留，这样就剔除了数据中的噪声。

小波分析的思想源于伸缩与平移的方法。设 $f(t)$ 是平方可积函数，记作 $f(t)\in L^2(R)$，则 $f(t)$ 的小波变换定义为：

$$W_f(a,b)=\int f(t)\frac{1}{\sqrt{a}}\psi(\frac{t-b}{a})dt \qquad (7)$$

式中的 a 为尺度因子，b 为平移因子，$\frac{1}{\sqrt{a}}\psi(\frac{t-b}{a})$ 是经过基本小波 $\psi(t)$ 平移和伸缩形成的函数系，这里 $\psi(t)$ 要满足如下允许性条件：

$$C_\psi=\int_0^\infty\frac{|\psi(\omega)|^2}{\omega}d\omega<\infty \qquad (8)$$

利用连续小波交换进行计算，计算量非常大，因此要考虑小波变换的快速算法。1988 年，Mallat 在构造正交小波基时提出多分辨率分析的概念，从空间概念形象的说明了小波的多分辨特性，并给出正交小波交换的快速算法，即 Mallat 算法。

小波分解的 Mallat 快速算法;

$$c_{jk}=\sum_m h(m-2k)c_{j-1,k} \qquad (9)$$

$$d_{jk}=\sum_m g(m-2k)d_{j-1,k} \qquad (10)$$

其中，$h(n),g(n)$ 分别对应低通和高通滤波器，上式表明，j 尺度上空间的尺度系数 d_{jk} 可由 $j-1$ 尺度系数 $c_{j-1,k}$ 经滤波器系数 $h(n),g(n)$ 进行加权和得到。

小波重构的 Mallat 快速算法：

$$c_{j-1,m}=\sum_k c_{jk}h(m-2k)+\sum_m d_{jk}g(m-2k) \qquad (11)$$

运用小波转换主要分为信号分解、小波细节系数加阈值、信号重建几个步骤，其中选取小波母函数和确定阈值为着重考虑的问题。

5.3.2. 基因信号去噪

本文将每个基因看做一个信号，基因在 62 个样本中的表现值看做不同时段的信号输入进行处理。

在信号处理中，噪声可以分为加法性噪声和乘法性噪声，大部分情况下都是加法性噪声[5]，根据这一特点，我们将基因表达数据在芯片制作、基因表达试验或者基因芯片表达图谱的采集转换过程中产生的噪音统一归纳为一种噪音 n，建立结肠癌基因

的信号含噪模型为：

$$y_i = x_i + Bn_i \quad (i = 1,2,3.....2000) \qquad (12)$$

式中 n_i 为零均值的高斯白噪声， y_i 为含噪信号， x_i 为需要提取的有用信号， B 为噪声方差。滤除噪声 n_i 的问题，可以认为是如何将期望值 x_i 从观测值 y_i 中恢复出来。

根据以上模型，结合小波变换的相关知识，本小组运用功能强大的 MATLAB 小波工具箱进行一维的基因信息降噪处理。主要处理过程可分为三部分：

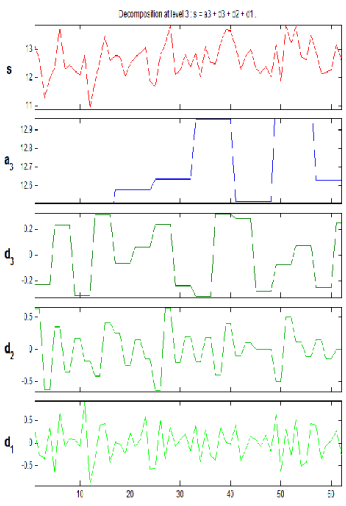
- 信号分解：选择小波基函数和分解的层次N，计算各层小波分解系数。
- 给小波细节系数加阈值：针对从1到N的每一层，分别选取阈值，并采用软阈值或硬阈值的方法进行阈值过滤。
- 信号重建：利用第N层近似值系数和1到N层的经过阈值过滤的细节系数对信号进行重建，得到的重建信号就是降噪后的信号。

在整个基因信息降噪的过程中，小波基函数的选取以及阈值的选取是最难确定的步骤，下面就按照前文提到的步骤对基因信号进行降噪处理：

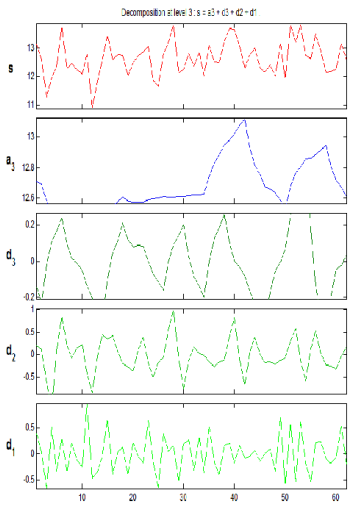
（1）信号分解

针对基因信号的无规律性，本文选择来自不同小波家族的正交小波基函数进行试验，包括来自Duabeehies家族的dbl到db7小波，来自Symlet家族的sym2小波，以及来自Coiflet家族的coif4小波。抽取基因号为H55933的基因数据作为试验信号。

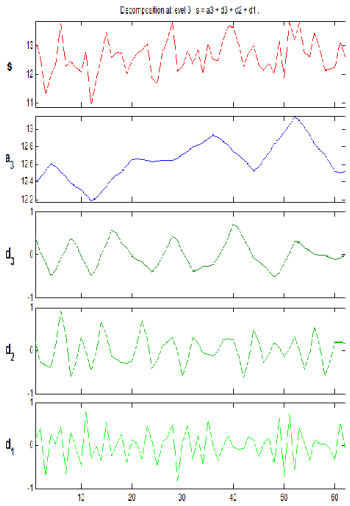
H55933基因运用不同基小波进行3层分解后如下图所示：



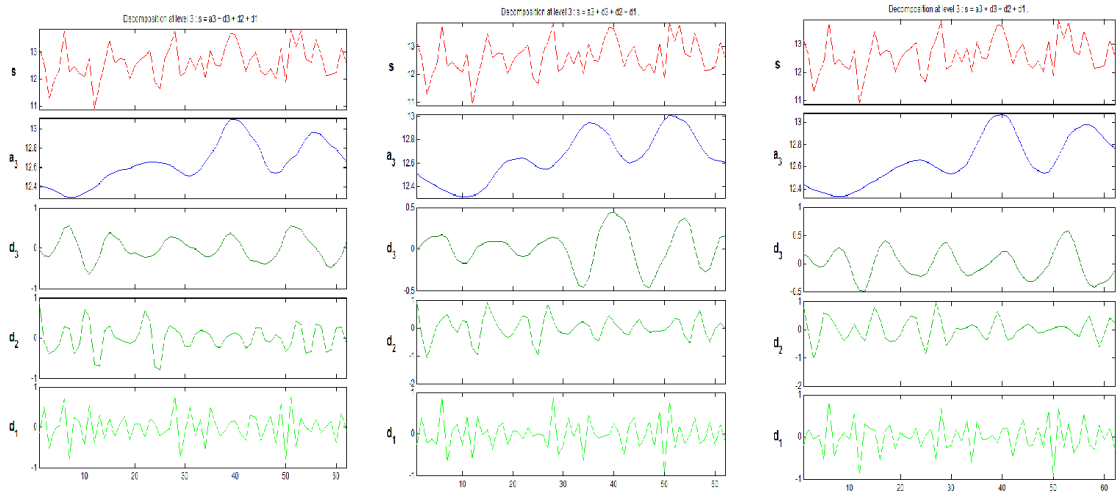
db1基小波三层分解



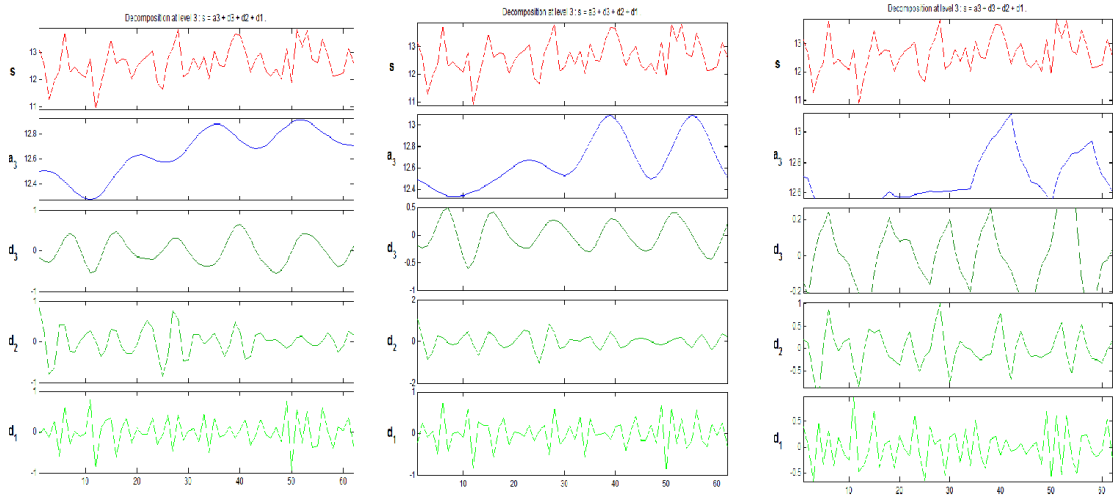
db2基小波三层分解



db3基小波三层分解



db4基小波三层分解　　db5基小波三层分解　　db6基小波三层分解



db7基小波三层分解　　coif4基小波三层分解　　sym2基小波三层分解

图 11　基因信号用不同基小波分解的三层函数图像

上图给出了用不同基小波进行分解所得到的基因分层函数图像。基因信号函数可以表示为：

$$s = a_3 + d_1 + d_2 + d_3 \qquad (13)$$

其中，s为原始信号，a_3表现为基因信号的低频数据，d_1、d_2、d_3为经过三层分解的高频基因信号。低频信号主要反映基因数据的近似部分，而高频基因信号则主要反映基因信号的细节部分。信号噪音主要分布在高频数据中，选择合适的阈值对三层信号进行处理理论上可以有效去除信号噪音。

（2）给小波细节系数加阈值

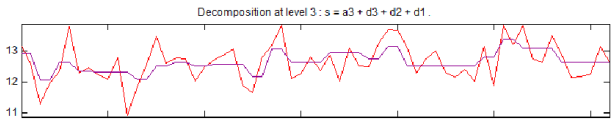
将分层好的数据根据合适的阈值进行过滤可以去除基因信号中的噪音。小波变换的阈值分为硬阈值和软阈值两种，硬阈值相对简单，但软阈值具有良好的数学特征表现。

本文比较软阈值中的固定阈值、自适应无偏似然阈值选择、启发式阈值、极大极小阈值以及能量阈值之后认为，基因表达数据中的噪声是潜在的、固有的系统随机噪声，其分布是未知的。选择传统的软阈值或者硬阈值方法都具有一定的随意性，难以确定其是否很好的反映信号本身的特征。而采用Rigorous阈值方法，在噪声分布未知的情况下，仍然可以较好地得到能量较小的系数作为噪声进行去噪处理，是一种相对合理的阈值方法。

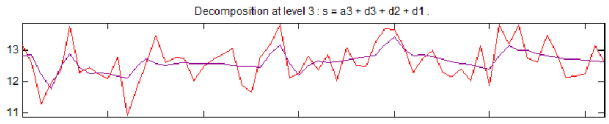
通过分析比较，确定选择Rigorous阈值作为小波阈值估计的方法。

（3）信号重建

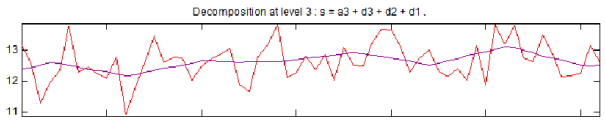
去噪之后的基因信号有第三次的近似信号与通过阈值过滤后的细节系数合并得到。下图给出几种小波基函数处理情况下原始H55933基因信号与去噪后的基因型号进行对比：



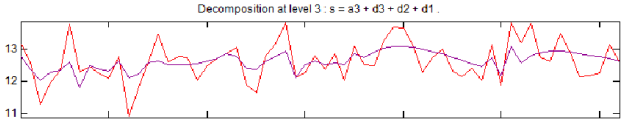
db1基小波下原始数据与去噪数据对比



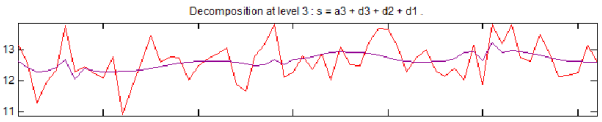
db2基小波下原始数据与去噪数据对比



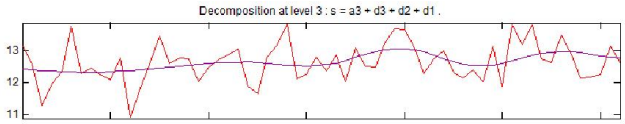
db3基小波下原始数据与去噪数据对比



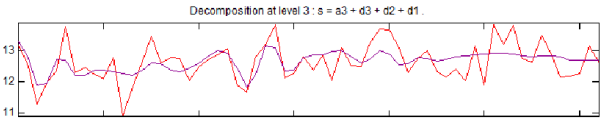
db4基小波下原始数据与去噪数据对比



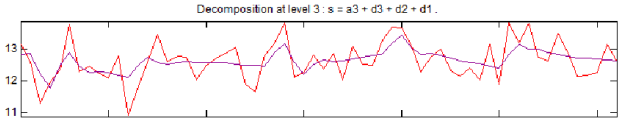
db5基小波下原始数据与去噪数据对比



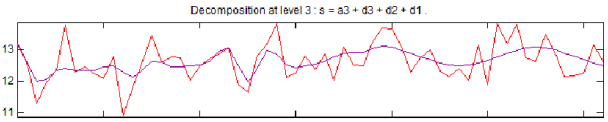
db6基小波下原始数据与去噪数据对比



db7基小波下原始数据与去噪数据对比



sym2基小波下原始数据与去噪数据对比



coif4基小波下原始数据与去噪数据对比

注：图中红色信号线为原始基因信号

紫色线为去噪处理重建后的基因信号

图 12 各种基小波下原始数据与去噪数据对比图

为了判断以上不同基小波处理得出的去噪基因信号哪个效果更好，我们采用“信噪比”这一参数进行判别。“信噪比”指标即：

$$d = \frac{\mu_o - \mu_i}{\sigma_o + \sigma_i} \qquad i = 1,2,3....9 \qquad (14)$$

式中 μ_o 、 σ_o 分别表示原始H55933基因数据的均值和标准差， μ_i 、 σ_i 表示运用不同小波基函数去噪后的H55933基因数据。图10中各小波函数在各时间点的去噪信噪比见下表：

表 7　信噪比对比表

	均值	标准差	信噪比
db1	12.65418	0.310128	0.00873905
db2	12.65359	0.280177	0.00837584
db3	12.64856	0.214534	0.00297819
db4	12.64857	0.278598	0.00276843
db5	12.64466	0.219436	-0.0017136
db6	12.65501	0.217848	0.01070518
db7	12.6578	0.304651	0.01272402
sym2	12.65359	0.280177	0.00837584
coif4	12.65535	0.291439	0.01021363

通过观察“信噪比”可以看出，通过db7小波基函数过滤的基因函数信噪比最高，表明这组基因信号通过处理噪音最小，所以将所有2000个基因都通过以上方法进行去噪处理。

需要指出的是，去噪过程并不是将存在噪音的基因直接去掉，而是将基因信号中的噪音因素去除，因此，去噪之后基因数量还是2000个不变。

除此之外，通过db7小波基函数处理基因表达谱图片可以得到以下结果：

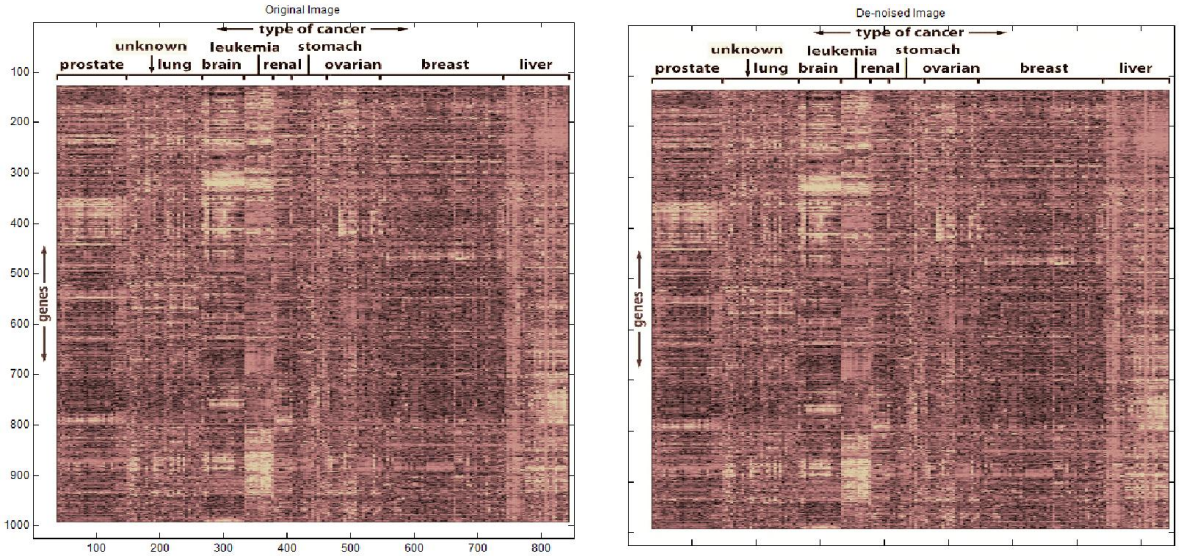


图 13　基因图谱通过小波转换去噪后的对比图

5.3.3. 过滤后基因信号进行分类

将两千个基因的基因样本数据按照以上方法进行小波去噪处理，得到的去噪数据请见附表2，将得到的新数据重新进行问题一中的GB综合指标分类以及问题二中的信息基因抽取，可以得到下文中对比结果。

（1）Gini 指标数据对比

将去噪后的2000个基因数据首先进行 *Gini* 指数的计算，下图为去噪后基因数据的 *Gini* 指数值升序排列分布曲线：

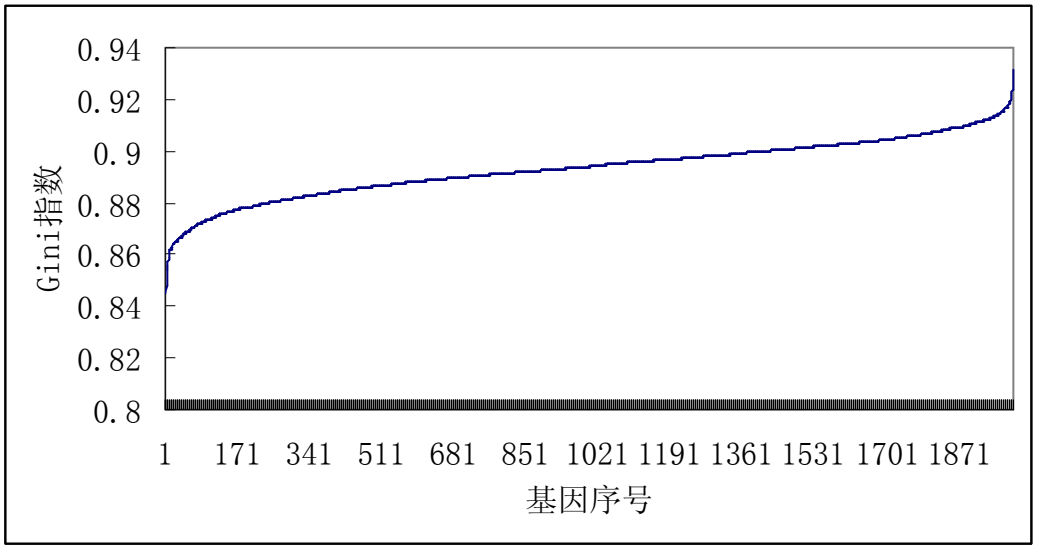


图 14　去噪后基因的 *Gini* 指数值升序排列分布曲线图

（2）Bhattacharyya 距离对比

将结肠癌基因数据重新计算Bhattacharyya距离，可得到如下结果：

表 8 去噪前后数据 Bhattacharyya 距离比较

Bhattacharyya 距离	基因个数		所占百分比	
	去噪前	去噪后	去噪前	去噪后
0～0.05	1571	1859	78.55%	92.95%
0.05～0.1	311	127	15.55%	6.35%
0.1～0.15	74	12	3.70%	0.60%
0.15～0.2	26	2	1.30%	0.10%
0.2～0.25	12	0	0.60%	0.00%
0.25～0.3	5	0	0.25%	0.00%
0.35～0.35	1	0	0.05%	0.00%

根据上表可以看出，通过去噪处理后，基因表达值的Bhattacharyya距离划分更加明显，92.95%都小于0.05，而0.2到0.35距离之间的距离都为零。通过下面的直方图可以更加清晰的看出去噪效果：

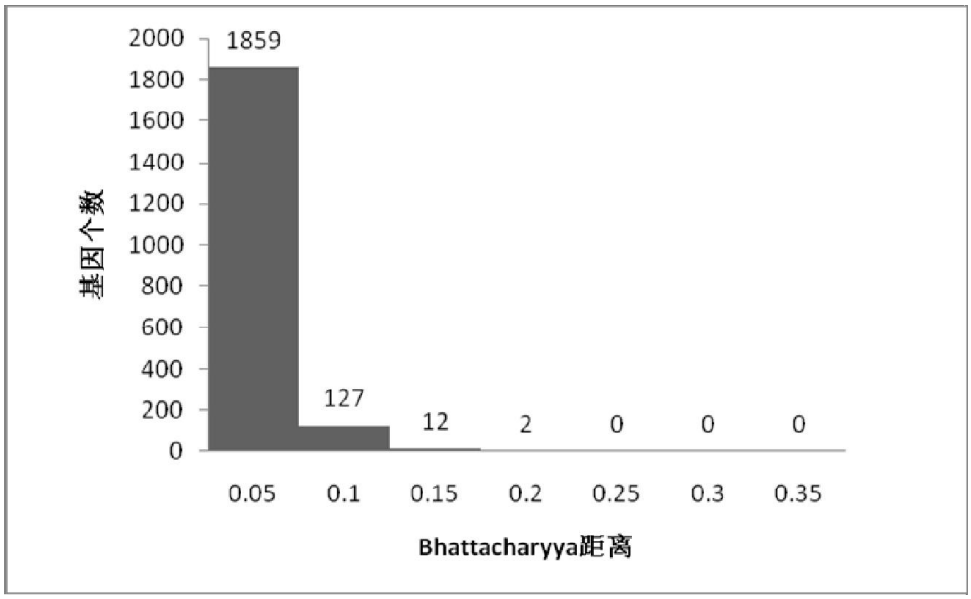


图 15 去噪后数据 Bhattacharyya 距离频数分布图

计算 *Gini* 指数和 Bhattacharyya 距离之后，通过设立合理的阈值来进行备用基因的选取。同样选取前 15%的基因作为备用基因，再选取两种方法的公共基因，选择的结果为 61 个公共基因，为：[H13281 L41268 J04794 J03569 R46528 R70939 M22632 R49416 L08187 U04641 D13665 H24310 X69550 T86473 R77220 X54942 R41873 R54467 T90350 U26312 H65355 T83368 M23410 T95018 R16156 R59583 T40674 X83299 X16663 D17390 X53586 U14973 D00762 H18490 T55558 T55558 M82919 R74066 U18299 R39130 R54097 R36977 U18920 X80497 U39360 R56399 T90549 H89092 M86737 D38521 R75893 T95063 L20422 T53396 M88108 X67325 R34698 T51261 M87772 R09138]。

可以看出，去噪后选取的信息基因更加精确，范围更小，说明去噪对于无效基因的筛除是有积极作用的。

（3）基因提取

选取好61个信息基因后，再利用这批基因进行分类特征基因子集的提取和分类预测效果的检验，具体步骤同第二题。结果如下：

表 9 去噪后各种基因子集的分错率统计

基因组合	F61	F55	F49	F45	F40	F36	F32	F29	F26	F23	F20	F18	F16
方法	GB 计算	两两冗余		BP 神经网络的 MIV 值分析									
错判数	3	3	3	3	5	3	6	3	3	5	3	3	3
基因组合	F14	F12	F10	F9	F8	F7	F6	F5	F4	F3	F2	F1	

方法	BP 神经网络的 MIV 值分析												
错判率	3	3	2	2	2	4	5	5	6	5	6	7	

上表的基因子集的错判数见图 14 折线图：

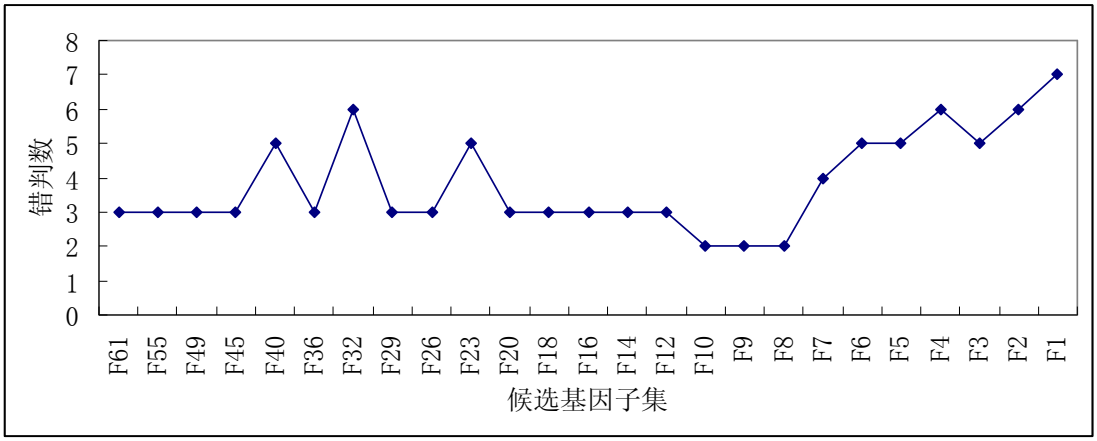


图 16 去噪后不同维数候选特征子集的分类能力

通过上表选取基因数量最少同时错判率最低基因组合，最终的基因子集为 F8，有以下基因：T62947， R84411，M82919，H43887，H08393，M26383，R36977，R87126。

从上面的结果可以看出，去噪后的最佳基因子集为 8 个基因，比去噪前要少了四个，同时去噪后的错判率更低了，说明去噪有明显的效果。

5.4. 问题四：信息基因求解的数学模型

5.4.1. 解题方法及假设

针对第四题，我们考虑运用基于 bayes 的层次聚类分析进行改进的方法求解。根据题目所给的数据，在问题四中我们提出如下假设：

1、 已经知道有 2 个信息基因（x_1 基因与 x_2 基因）与癌症关系密切，如题中所给的信息：大约 90%结肠癌在早期有 5 号染色体长臂 APC 基因的失活，有 40%~50%的 ras 相关基因突变。

2、 每个基因为一个个案，题中共有 2000 个个案，每个个案都为一个 62 维向量，即 $x_1=(x_{11},x_{12},\cdots,x_{1n})$， $x_2=(x_{21},x_{22},\cdots,x_{2n})$ （ $n=62$ ）。

5.4.2. 贝叶斯分析

根据给定的若干信息基因以及在对先验知识知之甚少的情况下，我们选择基于贝叶斯方法的 Q 型聚类算法，下面阐述该方法的基本思想和具体实现过程。

（1）Bayes 公式

设事件 $A_1,A_2,A_3\cdots\cdots A_k$ 构成互不相容的完备事件组，$\{P(A_j),j=1,2\cdots\cdots k\}$ 表示先验分布，由于事件 B 的发生，可以对事件 $A_1,A_2,A_3\cdots\cdots A_k$ 的发生提供新的信息，

$\{P(A_i/B), i=1,2\cdots\cdots k\}$ 表示后验分布，则概率论中的贝叶斯公式为

$$P(A_i/B)=\frac{P(B/A_i)P(A_i)}{\sum_{j=1}^{k}P(B/A_j)P(A_j)} \qquad (15)$$

引入离散型随机变量 θ，它的取值 $\theta_1,\theta_2,\theta_3\cdots\cdots\theta_k$，其中，$\theta_j=\theta(A_j)$，表示的是当 A_j 发生时，θ 的取值为 θ_j，先验分布 $\pi(\theta_j)=P(\theta=\theta_j)=P(A_j)$

B 是另外一个随机变量，定义一个随机变量 x，$x=x(B)$，即 $P(B/A_j)=P(x/\theta_j)=P(x/\theta=\theta_j)\,j=1,2\cdots\cdots k.$ 可以得到：

$$P(\theta_i/x)=P(\theta=\theta_i/x)=\frac{P(x/\theta_i)\pi(\theta_i)}{\sum_{j=1}^{k}p(x/\theta_j)\pi(\theta_j)}\,i=1,2\cdots\cdots k \qquad (16)$$

（2）基于 Bayes 方法的聚类分析

根据 Bayes 方法和 Q 型聚类的基本思想设计如下算法：

第一步：确定聚类中心数据值；

第二步：以聚类中心数据为聚类依据，根据先验信息假定出分布 $\pi(\theta)$，$\pi(\theta)$ 即为先验分布，并作为 Bayes 公式的先验概率。

第三步：调用聚类算法进行聚类（具体聚类过程请见下一小节）。

第四步：根据公式（1）计算出聚类后的后验概率。

第五步：用第四步得到的后验概率作为检验聚类结果的标准。若符合要求，则整个算法过程结束，否则则需要修改先验分布的部分参数或重新确定新的先验分布，知道所得的后验概率符合要求。

本题中，大约 90%结肠癌在早期有 5 号染色体长臂 APC 基因的失活，有 40%~50%的 ras 相关基因突变，在对先验知识知之甚少的情况下（符合本题的情况：信息基因识别中先验知识比较匮乏，且给出信息基因的先验概率）bayes 方法具有其他方法不可比拟的优点。

确定信息基因（x_1 基因与 x_2 基因）的初始聚类中心，记两个基因向量的任意凸线性组合为 $x=\sum_{i=1}^{62}\alpha_i wx_i\,(\sum_{i=1}^{62}\alpha_i=1,\alpha\geq 0)$，运用非线性规划方法在不同的基因表达水平之间寻找一致和妥协，即极小化可能的基因水平跟两个信息基因水平之间的总体偏差最小。非线性规划模型如下:

$$\begin{cases} \min \sum_{i=1}^{62} \left\| x - xi \right\|^2 \\ s.t. \sum_{i=1}^{62} \alpha_i = 1, \alpha_i \geq 0 \end{cases} \quad (17)$$

求解以上模型可得最满意的组合系数向量 $\alpha^* = (\alpha_1^*, \alpha_2^*)$，以及最满意的信息基因表达水平，也就是初始的类中心值 $x_c = (x_1^*, x_2^*, \cdots, x_n^*)(n = 62)$。

5.4.3. 调用聚类算法

上文中给出了基于贝叶斯分析的聚类分析整体思路，本节中给出利用重复逐层聚类进行基因信息分析的具体步骤。

（1）确定初始聚类中心

确定信息基因（x_1 基因与 x_2 基因）的初始聚类中心，记两个基因向量的任意凸线性组合为 $x = \sum_{i=1}^{62} \alpha_i w x_i (\sum_{i=1}^{62} \alpha_i = 1, \alpha \geq 0)$，运用非线性规划方法在不同的基因表达水平之间寻找一致和妥协，即极小化可能的基因水平跟两个信息基因水平之间的总体偏差最小。非线性规划模型如下:

$$\begin{cases} \min \sum_{i=1}^{62} \left\| x - xi \right\|^2 \\ s.t. \sum_{i=1}^{62} \alpha_i = 1, \alpha_i \geq 0 \end{cases} \quad (18)$$

求解以上模型可得最满意的组合系数向量 $\alpha^* = (\alpha_1^*, \alpha_2^*)$，以及最满意的信息基因表达水平，也就是初始的类中心值 $x_c = (x_1^*, x_2^*, \cdots, x_n^*)(n = 62)$。

（2）确定初始聚类种子及聚类中心点

指定步骤一中的 x_c 为初始聚类种子，再另指定 1 个聚类中心点。

（3）分类

根据距离最近原则进行分类。按照距离这 2 个类中心距离最近原则，把基因分配到各类中心所在的类中，形成一次迭代的 2 个分类。

（4）结果分析

对最终的分类结果进行分析，剔除与给定信息基因不在一类的个案，对于包含信

息基因的类中所有的个案重复前三个步骤，直到找到令人满意的的基因分类。

根据文献四，我们得知 T51023 和 Z50753 这两个基因为信息基因，得到聚类的初始类中心值x_c=（9.891838117　7.757628946　7.820468519　8.360936829
9.472687355　9.291795366　8.314878176　8.143527858　8.506667185
8.485060702　9.362765259　7.724640895　9.103498366　9.642246108
9.753726386　10.0139794　8.854795726　8.950332542　8.66024443
8.922453072　9.954906851　8.965655468　8.118224098　7.361392067
7.845776751　8.229829252　7.818288109　9.946772539　8.24182322
8.26654798　7.452612251　7.020921123　8.950165524　7.571927693
8.459064808　8.282485996　8.195837588　9.057345969　9.375895358
9.585535662　8.901487969　8.125193329　8.222188013　9.049927549
7.896504181　7.714192343　8.94529256　7.681532283　7.962284278
7.401246267　9.201890204　10.01782579　9.142266112　9.182786046
9.012470062　8.973423343　8.132722358　8.591634351　8.246531795
7.577382197　8.270686467　8.528622263）

6. 模型的评价与改进

6.1. 模型优点

（1）问题一的信息基因筛选中，将两种方法（*Gini* 指数和 Bhattacharyya 距离）进行结合，构建 GB 方法进行无关基因的去除，避免了现有方法的缺点和不足；

（2）在问题二的方法的选择上实现了创新，将相关性和 MIV 相结合，使整个候选特征基因筛选的过程更加有效，MIV 值的过程和灵敏度分析有着异曲同工之妙，在变量筛选的过程中有着很好的效果，值得推广；

（3）问题三中利用现有去噪技术最成熟的小波变换进行基因信息去噪处理，处理后的数据在基因分类过程中有更好的表现；

（4） 在第四文中引进贝叶斯知识来利用已知信息。

6.2. 模型缺点

（1）GB 综合指数的效果有一定的局限性

（2）MIV 值的引用虽具有创新性，但成熟性有待改进，值得进行深入研究和引用。

（3）问题三主要偏向于 MATLAB 工具箱应用，在噪音模型的建立方面有待完善。

（4）第四问中建立的利用已知信息的数学模型尚未完整，仅给出了基本思想及简单模型，未进行实践检验。

6.3. 模型的改进

现有的去除无关基因方法很多，但哪种方法最优尚无定论，需要大量的数据对现有的方法进行效果比较，总结和归纳；

MIV 值在其他领域进行自变量的筛选中应用很广泛，但在基因筛选中的使用有限，需要利用大量的数据和样本或不同的案例对该方法的有效性进行检验。

怎样利用好已知的临床信息对基因筛选很重要，将现有的自然科学定理和方法引用到该领域中具有很大的价值，这也会对医学的整体水平的提高提供思路。

参考文献

[1] T. R. Golub, et al. Monitoring and Class Prediction by Gene Expression， Science, Vol. 286, pp.531-537 ,1999.

[2] 刘全金，李颖新，阮晓钢 基于BP 网络灵敏度分析的肿瘤亚型分类特征基因选取[J].中国生物医学工程学报. Vol.27,No.5,pp.710-715. 2008,

[3] Guyon I , Weston J , Barnhill S , et al . Gene selection for cancer classification using support vector machines [J] . Machine Learning ,2000 ,46(13) :389 - 422.

[4] 刘全金，李颖新，朱云华等. 基于BP 神经网络的肿瘤特征基因选取[J].计算机工程与应用.,Vol.24.pp.184-186,216. 2005

[5] 宗孔德,胡广书.数字信号处理[M].北京:清华大学出版社, 1988.

[6] Z. Sun, P. Yang, Gene expression profiling on lung cancer Outcome Prediction: Present Clinical Value and Future Premise, Cancer Epidemiology Biomarkers & Prevention, Vol.15,No.11,pp. 2063-2068.2006

[7]李颖新,刘全金,阮晓钢，急性白血病的基因表达谱分析与亚型分类特征的鉴别，中国生物医学工程学报，Vol.24, No.2, pp.240-244(2005)

（相关数据请见文件夹Data File）

附件一　利用BP神经网络分类预测代码

```
a= xlsread('e:MIV.xls','sheet6','B2:M63');
P=a(1:40,:);
T= xlsread('e:MIV.xls','sheet6','N2:N41');
P=P';
T=T';
net=newff(minmax(P),[24,1],{'tansig','purelin'},'traingdm');
net.trainParam.show=50;
   net.trainParam.lr=0.05;
      net.trainParam.mc=0.9;
      net.trainParam.epochs=2000;
net=train(net,P,T);
R=a(41:62,:) ;
R=R';
Y=sim(net,R) ;
Yc=vec2ind(Y)
```

附件二　第二问MIV值筛选特征基因集

```
clear

  %设置网络输入输出值
  p= xlsread('e:MIV.xls','sheet4','B2:DL41');
  t= xlsread('e:MIV.xls','sheet4','DM2:DM41');
  t=t';

  %%变量筛选 MIV 算法的初步实现（增加或者减少自变量）


  [m,n]=size(p);
  yy_temp=p;

  %p_increase 为增加 10%的矩阵  p_decrease 为减少 10%的矩阵
  for i=1:n
        p=yy_temp;
        pX=p(:,i);
      pa=pX*1.4;
      p(:,i)=pa;
      aa=['p_increase' int2str(i) '=p'];
       eval(aa);
  end


  for i=1:n
      p=yy_temp;
      pX=p(:,i);
      pa=pX*0.6;
      p(:,i)=pa;
```

```
    aa=['p_decrease' int2str(i) '=p'];
    eval(aa);
end


  %%利用原始数据训练一个正确的神经网络
  nntwarn off;

  p=p';
  %BP 网络建立
  net=newff(minmax(p),[228,1],{'tansig','purelin'},'traingdm');
  %初始化 BP 网络
  net=init(net);
  %网络训练参数设置
  net.trainParam.show=50;
  net.trainParam.lr=0.05;
  net.trainParam.mc=0.9;
  net.trainParam.epochs=2000;

  %BP 网络训练
  net=train(net,p,t);

  %%变量筛选 MIV 算法的后续实现（差值计算）

  %转置后 sim

for i=1:n
    eval(['p_increase',num2str(i),'=transpose(p_increase',num2str(i),')'])
end
```

```
for i=1:n
    eval(['p_decrease',num2str(i),'=transpose(p_decrease',num2str(i),')'])
end

%result_in 为增加 10%后的输出  result_de 为减少 10%后的输出
for i=1:n
   eval(['result_in',num2str(i),'=sim(net,','p_increase',num2str(i),')'])
end

for i=1:n
    eval(['result_de',num2str(i),'=sim(net,','p_decrease',num2str(i),')'])
end

for i=1:n
    eval(['result_in',num2str(i),'=transpose(result_in',num2str(i),')'])
end

for i=1:n
    eval(['result_de',num2str(i),'=transpose(result_de',num2str(i),')'])
end
```

%%MIV_n 的值为各个项网络输出的 MIV 值 MIV 被认为是在神经网络中评价变量相关的最好指标之一，其符号代表相关的方向，绝对值大小代表影响的相对重要性。

```
for i=1:n
    IV=['result_in',num2str(i),'- result_de',num2str(i)];
    eval(['MIV_',num2str(i),'=mean( ',IV,')'])
end
```